DOCUMENT RESUME

ED 355 272 TM 019 624

AUTHOR Vansickle, Timothy R.
TITLE Work Keys: Developing a Usable Scale for Multi-Level, Criterion-Referenced Assessments.
PUB DATE Aug 92
NOTE 29p.; Version of a paper presented at the Annual Meeting of the American Psychological Association (100th, Washington, DC, August 14-18, 1992).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Criterion Referenced Tests; Employment Qualifications; High Schools; High School Students; Holistic Evaluation; Job Skills; Listening Comprehension Tests; *Mathematics Tests; Occupational Tests; *Reading Tests; *Scaling; Scores; *Test Construction; Test Reliability; Test Use; Test Validity; *Writing Tests
IDENTIFIERS American College Testing Program

ABSTRACT

The scaling of a new assessment is a significant undertaking. The scaling of a new assessment designed as a multiple-level, criterion-referenced assessment is even more so. A Guttman approach to scaling was used with the Work Keys selected-response assessments, Reading for Information and Applied Mathematics. Assessments in development in the Work Keys project are designed to aid in the communication of needed workplace skills to business persons, educators, and learners. Pretests were conducted with 5,741 high school students and adult employees who took the Reading for Information assessment, and 6,236 examinees who took the Applied Mathematics assessment. The classification rate of individuals into appropriate skill levels was very good, exceeding 95 percent. A similar procedure was developed for the holistic score scale for Listening and Writing (3,319 examinees). Research on the operational forms of these assessments must be conducted to determine the reliability of parallel forms and the validity of the instruments for various uses. However, the scaling procedures appear to be working well. Five tables contain study findings, and one figure illustrates the scoring procedure. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

[X] This document has been reproduced as received from the person or organization originating it.

[ ] Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

TIMOTHY R. VANSICKLE

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED355272

Work Keys: Developing a Usable Scale for Multi-level, Criterion-referenced Assessments

Timothy R. Vansickle

American College Testing

Revised Version of: In J. D. West (Chair), Work Keys: Supporting The Transition From School To Work. A Symposium presented at the One Hundredth Annual Meeting of the American Psychological Association, August 14, 1992, Washington, D. C.

TM019274

WORKING KEYS: GUTTMAN SCALING

BEST COPY AVAILABLE

Abstract

The scaling of a new assessment is a significant undertaking. The scaling of a new assessment designed as a multiple-level, criterion-referenced assessment is even more so. A Guttman approach to scaling was used with the Work Keys selected-response assessments, Reading for Information and Applied Mathematics. The classification rate of individuals into appropriate skill levels was very good, exceeding 95 percent. A similar procedure was developed for the holistic score scale used for the Work Keys constructed-response assessments, Listening and Writing. Research on the operational forms of these assessments needs to be conducted to determine the reliability of parallel forms and the validity of the instruments for various uses. However, the scaling procedures appear to be working well.

Work Keys: Developing a Usable Scale for Multi-level, Criterion-referenced Assessments

Since the release of *A Nation At Risk* in 1983, parents, educators, business leaders, and politicians have bemoaned the falling academic and workplace skills of American students, workers, and citizens. *America 2000: An Educational Strategy* released in 1991, stressed the need to improve our workforce skills to compete globally. In fact, the decline of workplace skills was noted before these reports were released. For example, achievement test scores for science began to decline in 1969. There have been, and will continue to be, many proposals for educational reform. These reforms stress back-to-basics curricula and/or technological skills.

In the past decade, the skills of students and workers have changed very little. Some people claim educators are at fault, others blame students, and still others cite testing methods as the reason for this lack of change. However, testing, even large-scale, group-administered, multiple-choice testing, is simply one form of assessment and not the cause of the lack of skills. Education relies heavily on testing or assessment, from teacher-developed tests to nationally normed tests of academic achievement. Although the use of tests is not the problem, teaching to a test can be a problem, especially when the test being taught to is narrow in scope of the overall domain of skill and knowledge.

Businesses also rely heavily on testing or assessment. The uses for tests in business range from screening applicants to determining promotion, advancement, and merit salary increases. Industrial and organizational psychology, the part of business most responsible for

assessment in corporate America, grew out of psychology's early success in measuring and describing individual differences. Today, industrial-organizational and human resource departments still rely on the measurement of individual differences with respect to abilities, aptitudes, values, interests, personality traits, and specific job skills. It is the latter category that has gained the most acceptance in business. While court rulings of the past decade dealing with test use and Equal Employment Opportunity Commission guidelines have much to do with the diminished use of some forms of testing in business, the use of tests for the purposes described above continue to be the dominant model.

Both education and business rely heavily on testing and assessment. However, a lack of communication exists between business, education, and the learner concerning what is needed to fill the gap in skills. A common language does not exist for these three stakeholders to communicate their respective needs about strengths and weaknesses of skills for individuals, programs, and/or occupations. If a common language is developed, communication between the concerned parties might be enhanced. This language should communicate an individual's strengths and weaknesses, not in relation to other individuals' performance, but in relation to some external criteria. One such criterion might be the type and level of skill needed for a specific job or class of related jobs (e.g., electrician, secretary, or bank teller). Another criterion may simply be how much of a subject domain an individual has mastered and/or what parts of the subject domain still need improvement.

If a suitable assessment of necessary skills were available, businesses could communicate to educators and learners the skills needed for success on the job, educators

could teach the needed skills, and learners would see the value in obtaining the necessary skills. Learners should be motivated to acquire the skills because of the link to future success in the workplace. Educators should be motivated to teach the skills that will make their students successful, and businesses will receive workers capable of successful completion of their assigned tasks. In addition, if the assessment of skills is completed early in a learner's educational process, learners will have the time to strengthen weak skill areas and acquire skills they lack.

The best approach for building this type of assessment is a criterion-referenced test (CRT) or assessment. Interest in CRTs has grown during the last decade or two (Hambleton and Rogers, 1991) because they assess an individual's performance with respect to a specified criterion rather than to the performance of other individuals. According to Popham (1978), "a criterion-referenced test is used to ascertain an individual's status with respect to a well-defined behavioral domain." Given the need for communication among business, education, and the learner, a CRT approach is the logical choice for assessing workplace skills and communicating information about the skills needed to all concerned. The assessments in development by Work Keys are of the CRT type and designed to aid in the communication of needed workplace skills to business persons, educators, and learners.

One of the many issues facing the Work Keys assessment program is to develop a scale for the criterion-referenced assessments that conveys meaning in a clear manner to anyone concerned about the strengths and weaknesses of individuals, programs, schools, training programs, etc. A scale is simply the ordering of things in some meaningful way

(Dunn-Rankin, 1983). As such, scaling involves the ordering of psychological objects or constructs. In addition to this ordering, large-scale assessment requires measuring a range of skills within a given domain. The Work Keys assessments have a broad range of skills within any given subject domain.

The issue of scaling is, therefore, a significant one. Louis Guttman (Stouffer, Guttman, Suchman, Lazarsfeld, Star, & Clausen, 1950) developed and described a unidimensional scaling procedure on which responses to items would place examinees in perfect order. This type of scaling has been labeled response centered (Crocker and Algina, 1986) because it simultaneously scales both the examinee and the items. However, this scaling procedure is deterministic in that it assumes no error in the items or examinees (Nunnally, 1978). That is, the probability of passing an item is 0 below the item's ability estimate and 1 beyond it. Regardless, the Guttman scaling procedure carries a great deal of meaning and is easily interpretable by users. The Work Keys assessments are intended to be meaningful tools to aid the learner, educator, and business person, and a Guttman scale offers a good method to report the scores of the multiple-level, criterion-referenced assessments.

The Guttman procedure involves simultaneously ordering examinees and items in an order of highest to lowest examinee score and easiest to most difficult item. Several indices can be computed based on the misfit of examinees and/or items. This misfit is essentially a type of error estimate reliability. Four indices are computed from the Guttman scaling procedure: coefficient of reproducibility, minimal marginal reproducibility, percent of

improvement, and the coefficient of scalability. These indices provide an estimate of the fit of the data to the Guttman model.

Method

Work Keys Pretests

The Work Keys system currently consists of three assessments which produce four scores. During the next two years, the Work Keys system will add six more assessments that will produce seven scores. Each assessment will be criterion-referenced with respect to the content domain it measures. The following discusses the development of the scales for the first three assessments: Reading for Information, Applied Mathematics, and Listening and Writing. The first two assessments are in a multiple-choice format. The latter is in a constructed-response format scored twice; once to determine how well the individual listened and retained/recorded information from an audiotaped message, and separately to measure the individual's writing ability.

The Work Keys assessments are contextualized by providing workplace situations, passages, problems, and messages for the examinee to respond to or solve. These situations and problems are similar to those one would find in a variety of occupations. Although the assessments are not specific to one particular occupation, some situations, problems, or items may represent one occupation more than another. However, no prior job-specific knowledge is required of the examinee. Someone who has completed a course in computer repair would not necessarily have an advantage when taking any of the Work Keys assessments. Within

any given assessment, the situations and problems represent many different types of occupations.

Each assessment was constructed with a number of levels. Each successive level is more difficult than the previous level. Difficulty was determined with respect to the cognitive load placed on the examinee in correctly responding to items within any given level. For example, the Applied Mathematics pretest contained five levels. The easiest level consisted of problems requiring application of simple arithmetic operations. The most difficult level consists of setting up multiple-step problems with unknowns and finding a solution. Given the design of the three pretested assessments, it appeared that a Guttman scaling would be feasible and that pretest data should provide a means of determining the fit of the data to the Guttman model.

Procedure

Each pretest required 90 minutes to administer. For both the Reading for Information and Applied Mathematics assessments, six pretest forms of 75 items were administered (total number of items per assessment was 450). For the Listening and Writing assessment, seven pretest forms of 12 recorded prompts were administered (total number of prompts was 72). A spiralled administration was used for the forms of both the Reading for Information and Applied Mathematics assessments. The Listening and Writing assessment is administered via audiotape and therefore, spiralling of forms was not possible. However, two of the 12 prompts in each form were anchor prompts (i.e., identical prompts) used in all seven forms. The anchor prompts provided a means of estimating and adjusting for any differences of the

intact groups taking the Listening and Writing assessment.

For both the Reading for Information and Applied Mathematics assessments, examinee responses were scored as either correct or incorrect. The Listening and Writing assessment was scored on a six-point scale of 0 to 5.

Pretest Sample

The Work Keys assessments were pretested in the spring of 1992 on a convenience sample of students and employees. Five Work Keys charter states volunteered to help pretest the assessments: Iowa, Ohio, Michigan, Tennessee, and Wisconsin. In most cases, an examinee took only one of the three assessments, and therefore, much of the discussion that follows will be by assessment.

The total sample size was 15,296 of which 5,741 examinees took the Reading for Information assessment, 6,236 examinees took the Applied Mathematics assessment, and 3,319 examinees took the Listening and Writing assessment. The sample consisted of approximately equal numbers of males and females, was 86 percent caucasian, and was 94 percent students regardless of the assessment. It should be noted that there was no intent to obtain a nationally representative sample because of the criterion-referenced nature of the assessments. Presented in Table 1 are the percentages of pretest examinees by various demographic categories for each of the three assessments.

Insert Table 1 about here

Results

Selected-Response Assessments

Within either the Reading for Information or the Applied Mathematics assessments, pretest items were judgmentally grouped by level of difficulty with the first set of items being least difficult and the last set of items being most difficult. These two assessments had five levels of difficulty each, with each level containing 15 items. It was hypothesized that if the items were working as planned, a scale score based on the level mastered would be the most informative way of reporting information back to the user. Furthermore, if this scaling procedure were working in a Guttman fashion, individuals who mastered the third set of items should also have mastered the first and second sets of items. Therefore, level scores would be assigned to an examinee by the most difficult contiguous (i.e., sequential) level mastered. What constituted mastery of a level? Based upon early prototype data and the input of several advisory panels, mastery of a level was tentatively set at 12 correct out of each set of 15 items (i.e., 80 percent correct).

In addition to the above level score, it was felt that information about the examinee's performance at the next, more difficult level would be important. This information could help describe how much of the next level was mastered and indicate what future steps the learner could take before the next administration of that assessment. Therefore, a partial score was devised to be the proportion of items answered correctly towards mastery of the next, more difficult level. An examinee might obtain a score of 3.5, which indicates the examinee had mastered the first three levels of a skill and was halfway to mastering the next

level of that skill. When provided with information about the skills in each level, the educator, business person, and learner would know what needs to be studied, practiced, or learned to achieve a score of 4.0 or above.

The Work Keys selected-response assessments were assigned a beginning level value of 3. For example, those individuals who answered 12 or more correct out of the first 15 items would receive a score of 3. The starting value of 3 was chosen because Work Keys was designed to begin measuring skills at a point where businesses would most likely be comfortable setting a minimum requirement. This starting value allows the development of levels below the beginning level for diagnostic or special use. Therefore, the range of scores would be 3.0 to 7.0.

Contiguity Analysis

Presented in Table 2 are contingency data comparing the most difficult contiguous level mastered with the most difficult level mastered regardless of contiguity based on total number of examinees for the Reading for Information assessment. The numbers in the diagonal boxes indicate the frequency of consistently classified examinees. The numbers below the diagonal indicate the frequency of inconsistently classified examinees. At the bottom of the table is the total number and the percentage of inconsistently classified examinees. In this instance, the total number inconsistently classified is 267 (4.7 percent).

Insert Table 2 about here

Similar results were obtained for the Applied Mathematics assessment. Here again, the total number of inconsistently classified examinees is low, 197 (3.2 percent). This type of contiguity information is one form of describing the reliability of the Work Keys assessments and is based on the score scale, not the items.

For both the Reading for Information and the Applied Mathematics assessments, individuals were consistently classified into their skill levels more than 95 percent of the time. This classification rate is very impressive considering pretest items were used and no misfitting items were removed for these analyses.

Guttman Scaling Analyses

The item-by-person Guttman procedure begins with item data and total scores. Table 3 contains the Guttman indices for each of the pretest forms from both the Reading for Information and Applied Mathematics assessments. These indices are computed from the item response data for all of the pretested items. Recall that the Guttman procedure scales (i.e., orders) items and individuals simultaneously. Two Guttman scale procedures were completed for each assessment; once using classical p-values and total test scores, and once using IRT parameters and ability estimates. These two analyses produced almost identical results. Therefore, presented in Table 3 are the results from the classical true score analysis. It should be noted that no misfitting items or individuals were removed from these analyses as is normally done in Guttman scaling. The values for the coefficient of reproducibility (CR) and coefficient of scalability (CS) are the most informative in determining the fit of the data to the Guttman model. For each of the assessments, the values approached or exceed

the critical values (i.e., CR $\geq$ .90 and CS $\geq$ .60). When misfitting items and individuals were removed (i.e., 5 or 6 items and 10 to 20 individuals per form) all the within form indices (i.e., CR and CS) exceeded the critical values.

Insert Table 3 about here

The results for the selected-response analyses indicate that the scale developed is usable and conveys meaning about the examinee's skills. It also appears that the data collected fits the Guttman model. Furthermore, the scale classifies individuals into skill levels with a great deal of consistency.

Constructed-Response Assessments

The constructed-response assessments consist of a set of audiotaped stimuli. The examinee responds to the stimuli by writing a message, paragraph, or short correspondence. The construction of this assessment was similar to that of the two assessments described previously. Each set of audio prompts was more difficult than the previous set of prompts. The stimuli were arranged into four levels with three audio prompts in each level. The difficulty of a prompt was based on the amount of information it contained. This varied from 7 pieces of information in the least difficult level to 16 pieces of information in the most difficult. The written responses of an examinee were scored twice, once for listening and once for writing. Both were scored with a holistic scoring procedure that had a scale of 0 to 5. Each assessment had its own descriptions and exemplars for each of the six score

points. Scorers were trained for specific pretest forms and scored only the listening or the writing but not both.

It was hypothesized that as the number of pieces of information in a prompt increased, the cognitive load increased, and therefore, should affect the examinee's performance. In other words, those with lower skills would have more difficulty with the prompts at the more difficult levels. It was also thought that this would be more true of the listening score than of the writing score since the writing score was based on how well the response was written and not on how much of the information was recorded accurately in the response.

Presented in Tables 4 and 5 are the contingency data from the Listening and the Writing assessments. Mastery of a level was determined by using the average score for the prompts within the level. Two separate cutoffs were used (i.e., 4.0 and 3.3) for these analyses. The tables contain the data for the cutoff set at 4.0. As shown in the tables, the consistency of classification is low. In fact, the error rate exceeds 20 percent for both assessments. It was obvious from these analyses that the level-based approach to developing a scale, which had worked so well with the selected-response assessments, was not a viable procedure. It was expected that this would occur for the writing assessment but not for the listening assessment. There are several different possible reasons for the above results: scorers use the middle of the holistic score scale, the convenience sample contains individuals with very similar skill levels, and/or listening skills are similar to writing skills and do not change with respect to complexity of the stimuli.

Insert Tables 4 and 5 about here

The results of the contiguity analysis suggest that the score scale could be based on the holistic score scale and associated descriptions and exemplars. Several procedures were developed using the holistic scale that would mimic the Guttman scaling used in the selected-response assessments.

A procedure was developed where the scores for the prompts would be tallied by the six holistic score points. To illustrate the procedure, assume the scores for an examinee's 12 prompts are 3, 4, 3, 4, 3, 4, 3, 3, 3, 2, 3, and 2 (see Figure 1). The tally procedure would be conducted in the following manner. First, tally the number of scores that are equal to, or greater than, the score category 0 (i.e., all 12). Second, repeat the first step for each of the remaining score categories. This yields the follow tallies (counts) per score category: 0, 12; 1, 12; 2, 12; 3, 10; 4, 3; and 5, 0. If a tally cutoff of 75 percent of the total number of prompts was set, then the integer part of the score would be 3 (i.e., 9/12 = .75; score point 3 is the largest score point that has a tally of nine or more). That is, this examinee performs consistently at the score point 3. The examinee also has received a score of 4, four times. Therefore, the decimal portion of the score is .444 (i.e., 4/9 = .444; 9 prompts considered mastery). The examinee's score would be 3.4 rounded to one decimal point. This procedure provides an interpretation scheme similar to that of the selected-response assessments. The examinee would know what skills need improvement to obtain a score of 4.0 or better. The

relationship (i.e., Pearson Product-Moment Correlation) between writing and listening scores was .52. This would indicate that the two assessments are measuring different skills as only 27 percent of the variance of one skill score is explained by the skill score of the other assessment.

The data for the seven pretest forms were further analyzed using the SPSS-X Reliability procedure. Since the derived score is not a linear transformation of the prompt scores, it would not be appropriate for use with the SPSS-X procedure. Therefore, internal consistency and a strict parallel unbiased reliability were computed using the SPSS-X Reliability procedure with the total sum of prompts for each examinee as the score. Coefficient alpha ranged from .89 to .92 for writing and .74 to .81 for listening.

Discussion

It would appear that for the dichotomously scored selected-response assessments, the data fit the Guttman scale model well. The scale allows the user to interpret an examinee's skills in terms of strengths and weaknesses. Furthermore, this type of score scale provides information to examinees that helps them determine what steps need to be taken to improve skills. However, the constructed-responses for Listening did not fall neatly into place with respect to a Guttman scale model. Therefore, the procedure adopted for reporting scores for Listening and Writing provides the same type of interpretive information to the examinee as the scaling procedure for the Reading for Information and Applied Mathematics assessments. The scales developed using the pretest data will, of course, be cross-validated as the first operational data is processed during the fall of 1992 and the spring of 1993. In addition, the

scaling procedures will be extended to the new assessments being developed over the next two years.

Overall, the Work Keys program, as designed, appears to have a solid foundation and should easily support a variety of uses. The score scales should be much more meaningful, to examinees and decision-makers, than the traditional standard score or percentile ranking. As more data are collected, the foundational aspects can be cross-validated and expanded.

References

Crocker, L. & Algina, J. (1986). Introduction to classical and modern test theory. New York, NY: Holt Rinehart Winston Publishing.

Dunn-Rankin, P. (1983). Scaling methods. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hambleton, R. K., & Rogers, H. J. (1991). Advances in criterion-referenced measurement. In R. K. Hambleton, & J. N. Zaal (Eds.), Advances in Educational and Psychological Testing.

National Commission on Excellence in Education (1983). A nation at risk: The imperative of educational reform. Washington, DC: U.S. Government Printing Office.

Nunnally, J. C. (1978). Psychometric theory (2nd ed.) New York, NY: McGraw-Hill.

Popham, W. J. (1978). Criterion-referenced measurement. Englewood Cliffs, NJ: Prentice-Hall.

Secretary's Commission on Achieving Necessary Skills (SCANS) (1991). America 2000: An education strategy. Washington, DC: U.S. Government Printing Office.

Stouffer, S. A., Guttman, L., Suchman, E. A., Lazarsfeld, P. F., Star, S. A., & Clausen, J. A. (1950). Measurement and prediction. Princeton, NJ: Princeton University Press.

Table 1

Percentage of Pretest Examinees in Various Demographic Categories by Work Keys Assessment

Demographic Category	Reading For Information	Applied Mathematics	Listening and Writing
Gender			
Male	49.5	49.5	48.5
Female	50.5	50.5	51.5
Race/Ethnicity			
African-American/Black	5.2	4.8	4.5
Caucasian/White	85.1	86.7	87.2
All Other Combined	9.7	8.5	8.3
Education Level			
9	40.4	38.4	27.9
10	7.3	9.4	5.8
11	13.1	13.3	18.3
12	33.2	33.3	40.0
High School Graduate +	6.0	5.6	8.0
Educational Program			
General Education	45.8	51.0	40.3
Vocational/Technical	22.4	17.2	30.5
College Preparatory	30.0	30.0	27.3
Other	1.8	1.8	1.9
School/Work Status			
Student	97.0	97.5	94.5
Employee	3.0	2.5	5.5

Table 2

Contingency Table for Reading for Information for Total Pretest Sample

		Highest Contiguous Level Achieved							
		Below Level 3	Level 3	Level 4	Level 5	Level 6	Level 7	Row Totals	
	Below Level 3	734						734	12.8%
	Level 3		1,547					1,547	26.9%
Highest	Level 4	55		1,699				1,754	30.6%
Level	Level 5	14	128		1,377			1,519	26.5%
Mastered	Level 6		5	8		98		111	1.9%
	Level 7		1	7	49		19	76	1.3%
	Column Totals	803	1,681	1,714	1,426	98	19	5,741	100.0%
		14.0%	29.3%	29.9%	24.8%	1.7%	0.3%	100.0%	
	Total Errors	267							
		4.7%							

Table 3

Guttman Coefficients for the Reading for Information and Applied Mathematics Assessment Pretests

Assessment Index	Pretest Forms 02	03	04	05	06	07
Reading For Information						
Coefficient of Reproducibility	0.89	0.89	0.90	0.89	0.89	0.90
Percentage of Improvement	0.16	0.16	0.15	0.16	0.16	0.15
Coefficient of Scalability	0.60	0.59	0.60	0.59	0.61	0.59
Minimal Marginal Reproducibility	0.73	0.73	0.75	0.74	0.73	0.75
Applied Mathematics						
Coefficient of Reproducibility	0.88	0.88	0.88	0.88	0.87	0.88
Percentage of Improvement	0.16	0.15	0.16	0.17	0.16	0.16
Coefficient of Scalability	0.57	0.57	0.57	0.58	0.56	0.57
Minimal Marginal Reproducibility	0.72	0.73	0.72	0.71	0.72	0.73

Table 4

Contingency Table for Listening for Total Pretest Sample

		Highest Contiguous Level Achieved						
		Below Level 3	Level 3	Level 4	Level 5	Level 6	Row Totals	
	Below Level 3	1,115					1,115	36.9%
Highest	Level 3		637				637	21.1%
Level	Level 4	188		219			407	13.5%
Mastered	Level 5	172	180		133		485	16.1%
	Level 6	106	131	53		84	374	12.4%
	Column Totals	1,581	948	272	133	84	3,018	100.0%
		52.4%	31.4%	9.0%	4.4%	2.8%	100.0%	
	Total Errors	830						
		27.5%						

Table 5

Contingency Table for Writing for Total Pretest Sample

		Highest Contiguous Level Achieved						
		Below Level 3	Level 3	Level 4	Level 5	Level 6	Row Totals	
	Below Level 3	1,764					1,764	58.4%
Highest	Level 3		240				240	8.0%
Level	Level 4	156		130			286	9.5%
Mastered	Level 5	92	67		64		223	7.4%
	Level 6	171	85	76		173	505	16.7%
	Column Totals	2,183	392	206	64	173	3,018	100.0%
		72.3%	13.0%	6.8%	2.1%	5.7%	100.0%	
	Total Errors	647						
		21.4%						

Figure 1.
Holistic Scoring Procedure Based On Majority of Prompts
At or Above a Score Point.